An (very brief) introduction to Deep Learning

for Computer Vision

Prof. Dr. Laura Leal-Taixé

www.dvl.in.tum.de

Technische Universität München

Computer Vision

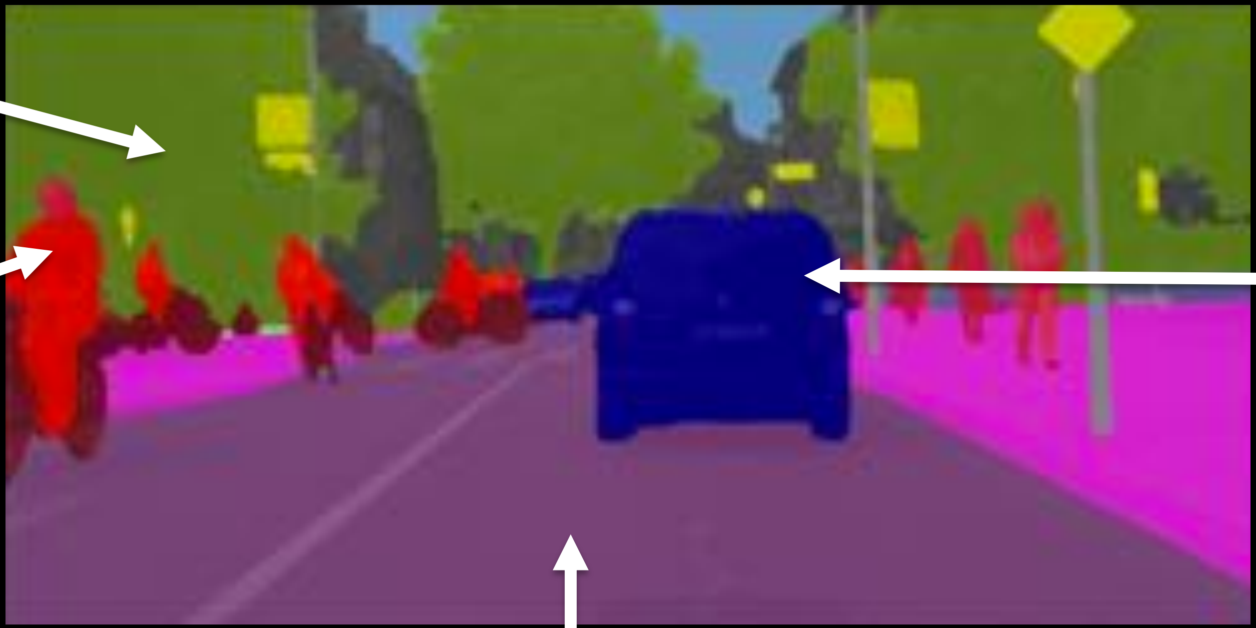
Give eyes to a computer

Computer Vision

Understand every pixel of an image

Computer Vision

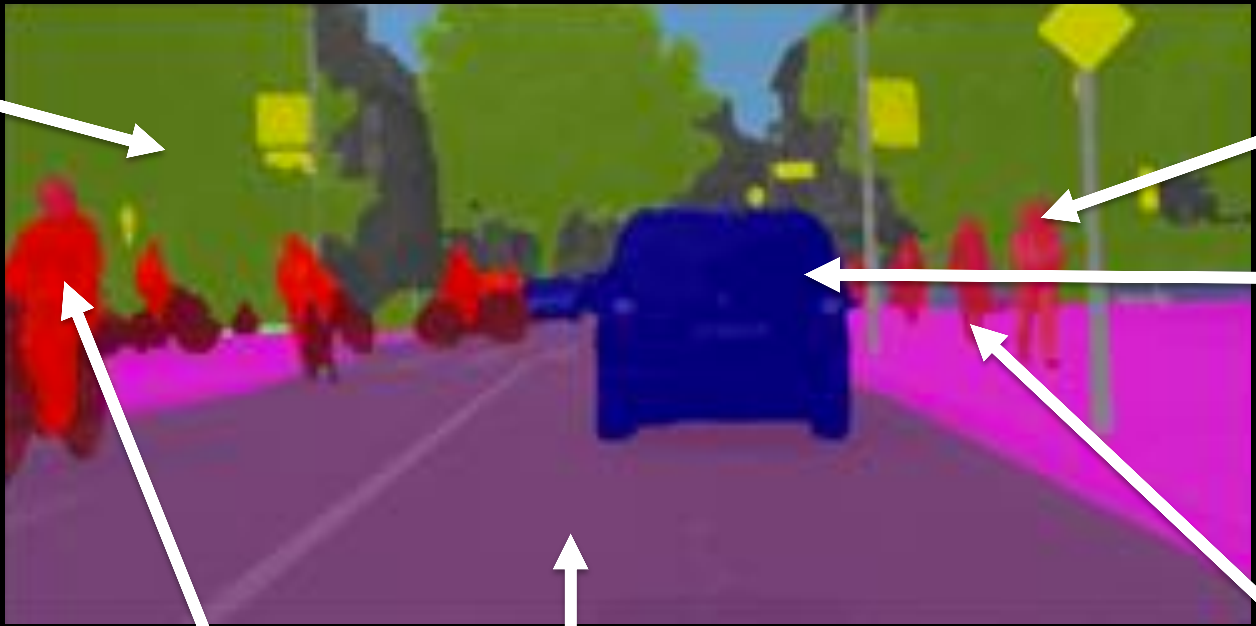
Understand every pixel of an image



Semantic segmentation

Computer Vision

Understand every pixel of an image



Instance-based segmentation

Semantic segmentation

Computer Vision

Understand every pixel of a video

Multiple object tracking

Instance-based segmentation

Semantic segmentation

Dynamic Scene Understanding

Understand every pixel of a video

Multiple object tracking

Instance-based segmentation

Semantic segmentation

Artificial Intelligence

Computational models

Data,
examples

Perform a given task on
new unseen data

LEARNING,
TRAINING

Artificial Intelligence

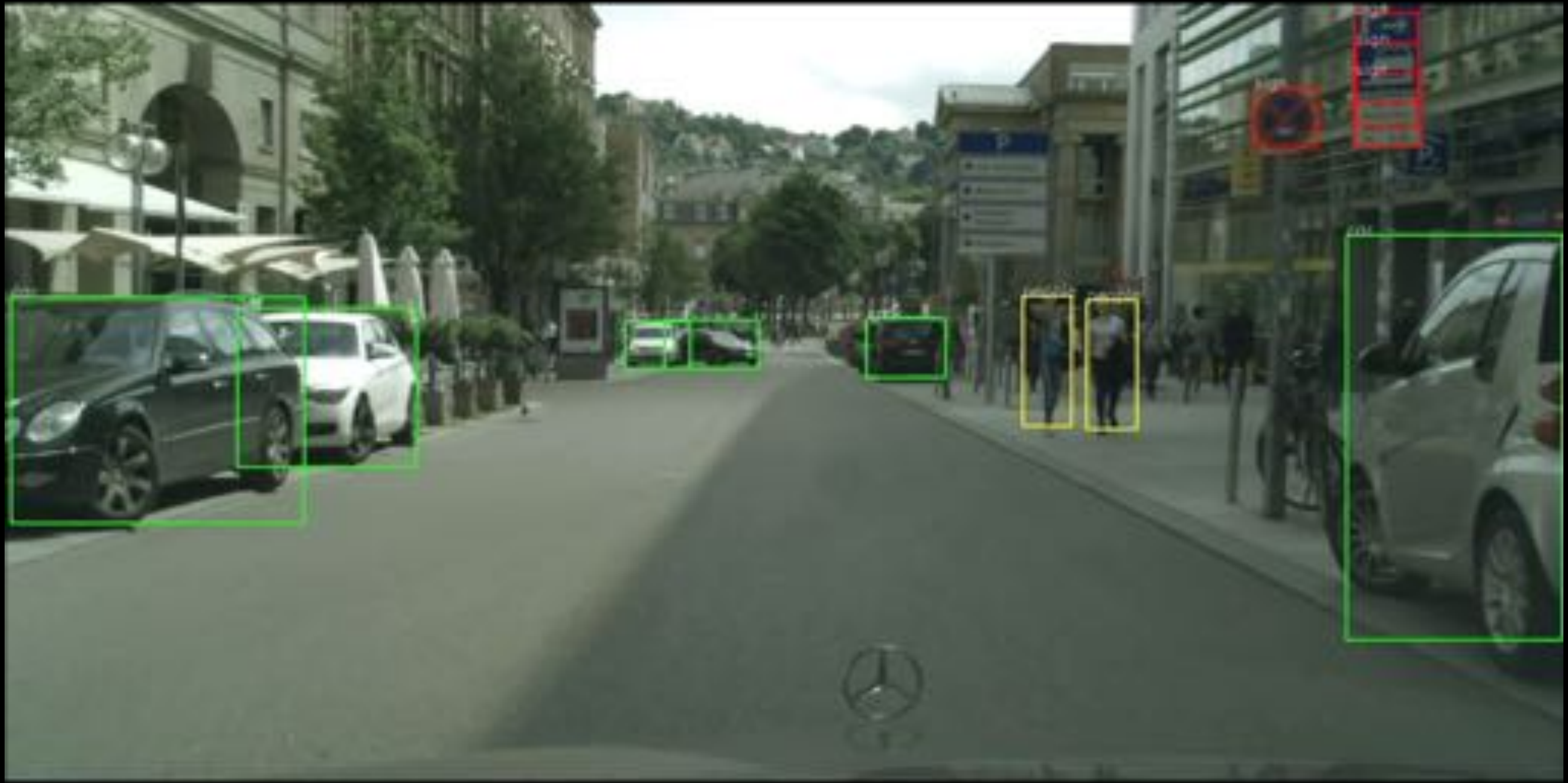
Computational models

Data,
examples

Perform a given task on
new unseen data

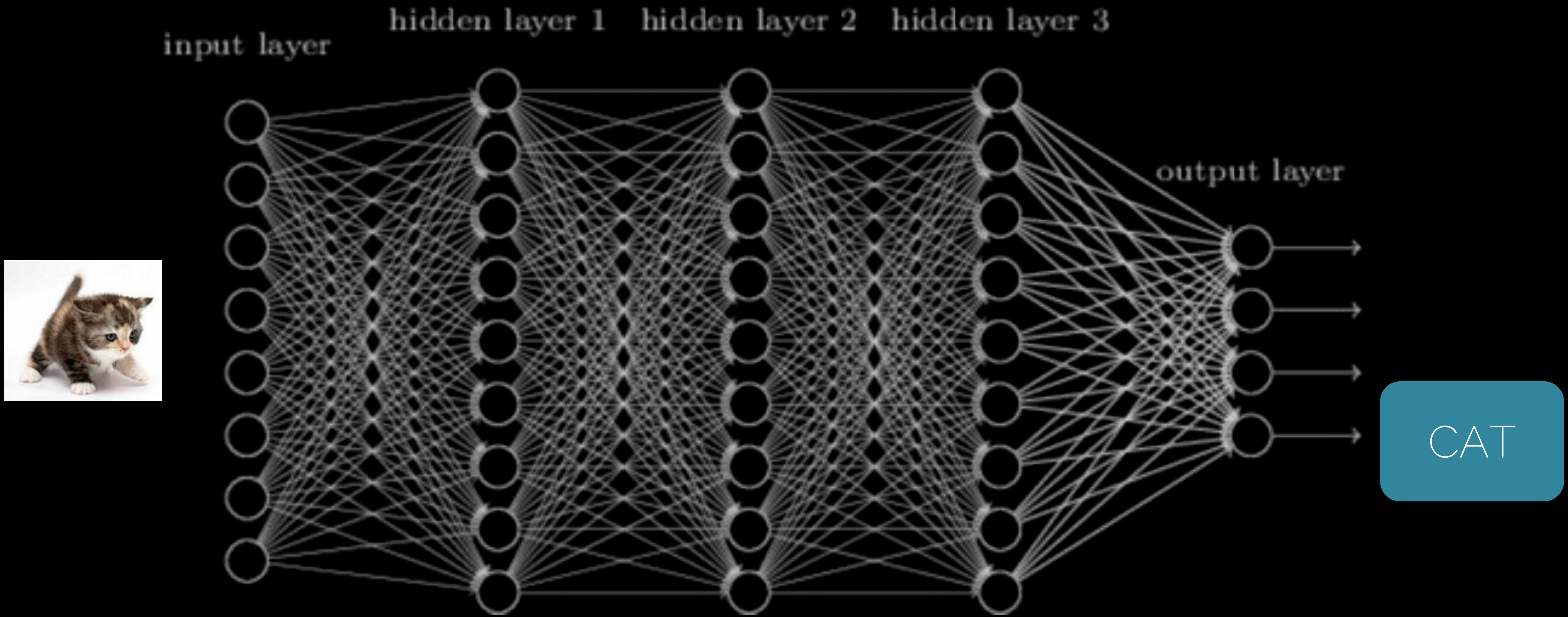
Expert eye

AI nowadays

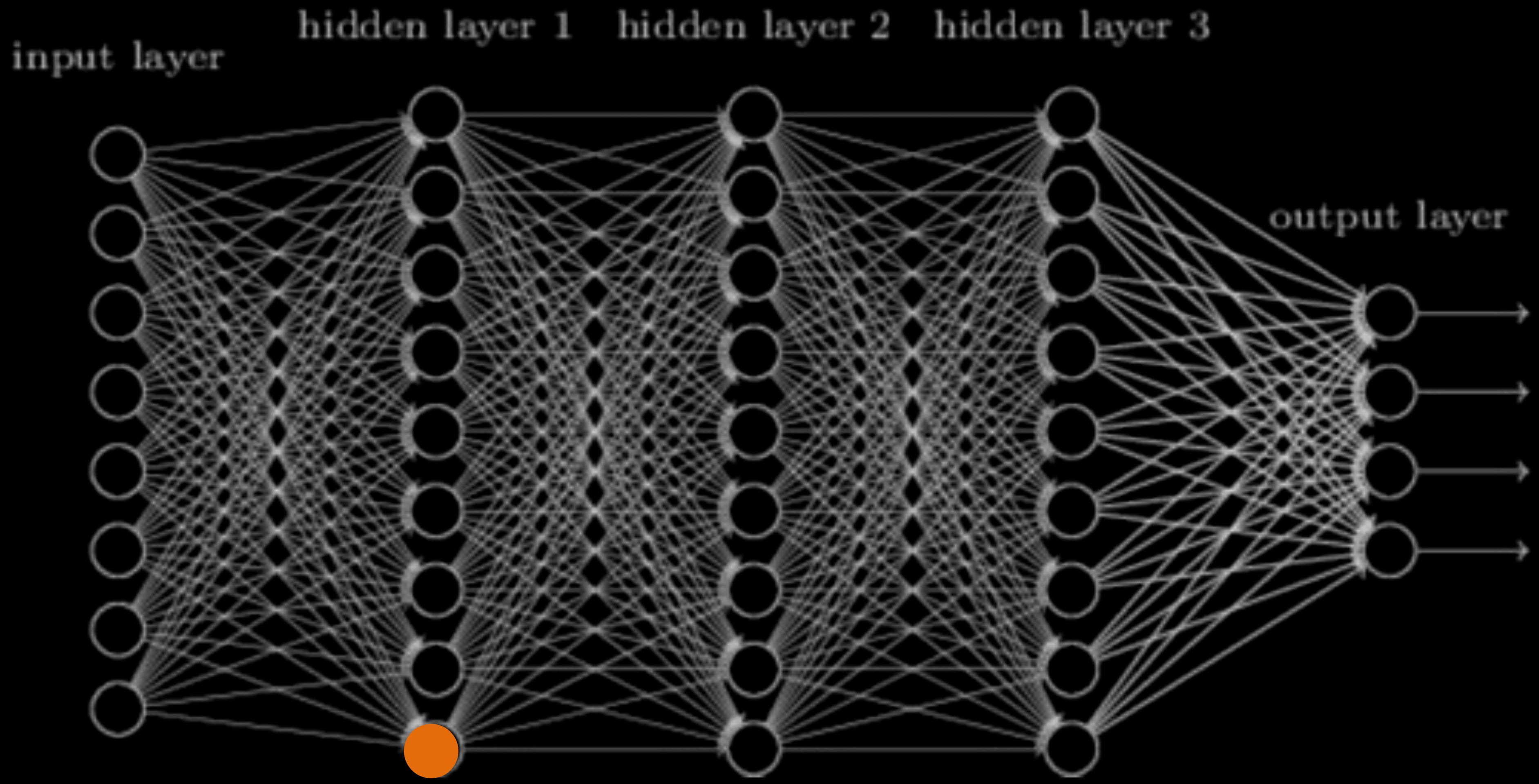


Self-driving cars

What is Deep Learning, really?

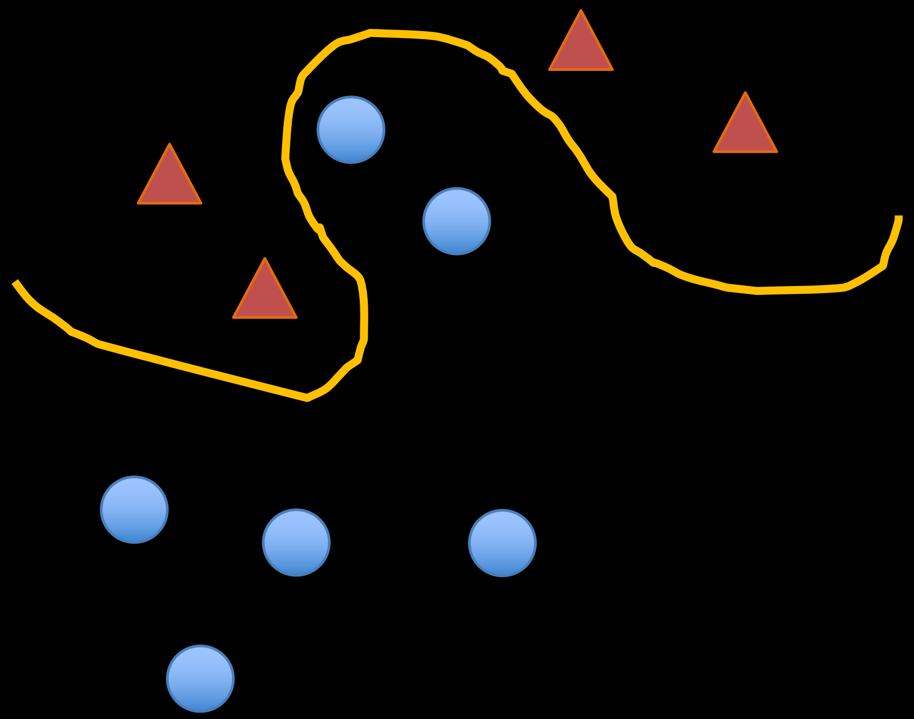


What is Deep Learning, really?



Each node is a small classifier

What is Deep Learning, really?



Each node is a small classifier

Each classifier makes tiny decision

Furry

Has two eyes

Has a tail

Has paws

Has two ears

These are learned from data!

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6		

$$5 \cdot 3 + (-1) \cdot 3 + (-1) \cdot 2 + (-1) \cdot 0 + (-1) \cdot 4 =$$
$$15 - 9 = 6$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	

$$5 \cdot 2 + (-1) \cdot 2 + (-1) \cdot 1 + (-1) \cdot 3 + (-1) \cdot 3 =$$
$$10 - 9 = 1$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

$$5 \cdot 1 + (-1) \cdot (-5) + (-1) \cdot (-3) + (-1) \cdot 3 + (-1) \cdot 2 =$$
$$5 + 3 = 1$$

Output 3x3

6	1	8

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7		

$$5 \cdot 0 + (-1) \cdot 3 + (-1) \cdot 0 + (-1) \cdot 1 + (-1) \cdot 3 =$$
$$0 - 7 = -7$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7	9	

$$5 \cdot 3 + (-1) \cdot 2 + (-1) \cdot 3 + (-1) \cdot 1 + (-1) \cdot 0 =$$
$$15 - 6 = 9$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7	9	2

$$5 \cdot 3 + (-1) \cdot 1 + (-1) \cdot 5 + (-1) \cdot 4 + (-1) \cdot 3 =$$
$$15 - 13 = 2$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7	9	2
-5		

$$5 \cdot 0 + (-1) \cdot 0 + (-1) \cdot 1 + (-1) \cdot 6$$
$$+ (-1) \cdot (-2) = -5$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7	9	2
-5	-9	

$$5 \cdot 1 + (-1) \cdot 3 + (-1) \cdot 4 + (-1) \cdot 7 + (-1) \cdot 0 =$$
$$5 - 14 = -9$$

Convolutions on Images

Image 5x5

-5	3	2	-5	3
4	3	2	1	-3
1	0	3	3	5
-2	0	1	4	4
5	6	7	9	-1

Kernel 3x3

0	-1	0
-1	5	-1
0	-1	0

Output 3x3

6	1	8
-7	9	2
-5	-9	3

$$5 \cdot 4 + (-1) \cdot 3 + (-1) \cdot 4 + (-1) \cdot 9 + (-1) \cdot 1 =$$
$$20 - 17 = 3$$

Image filters

- Each kernel gives us a different image filter

Input

Edge detection

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Box mean

$$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Sharpen

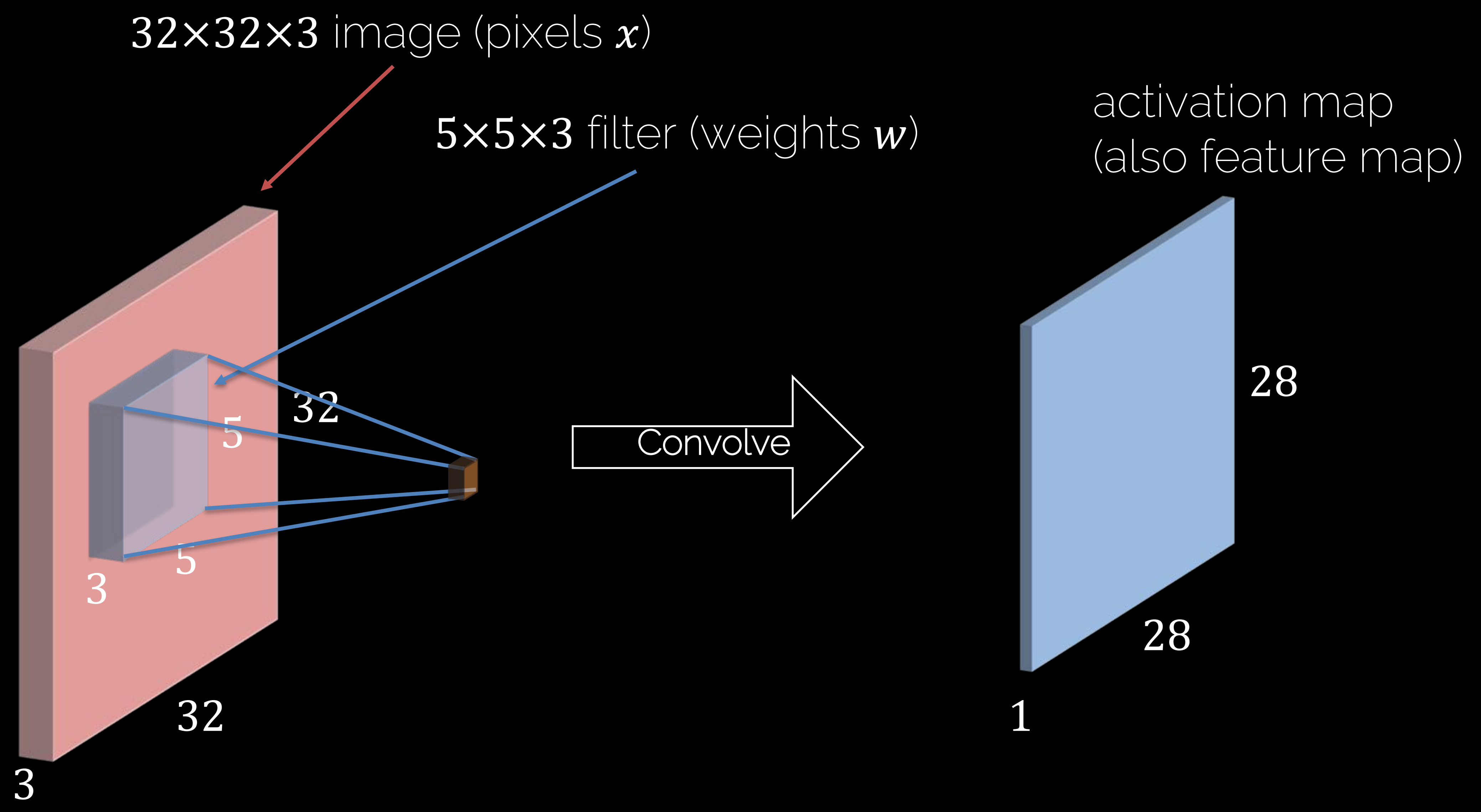
$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

Gaussian blur

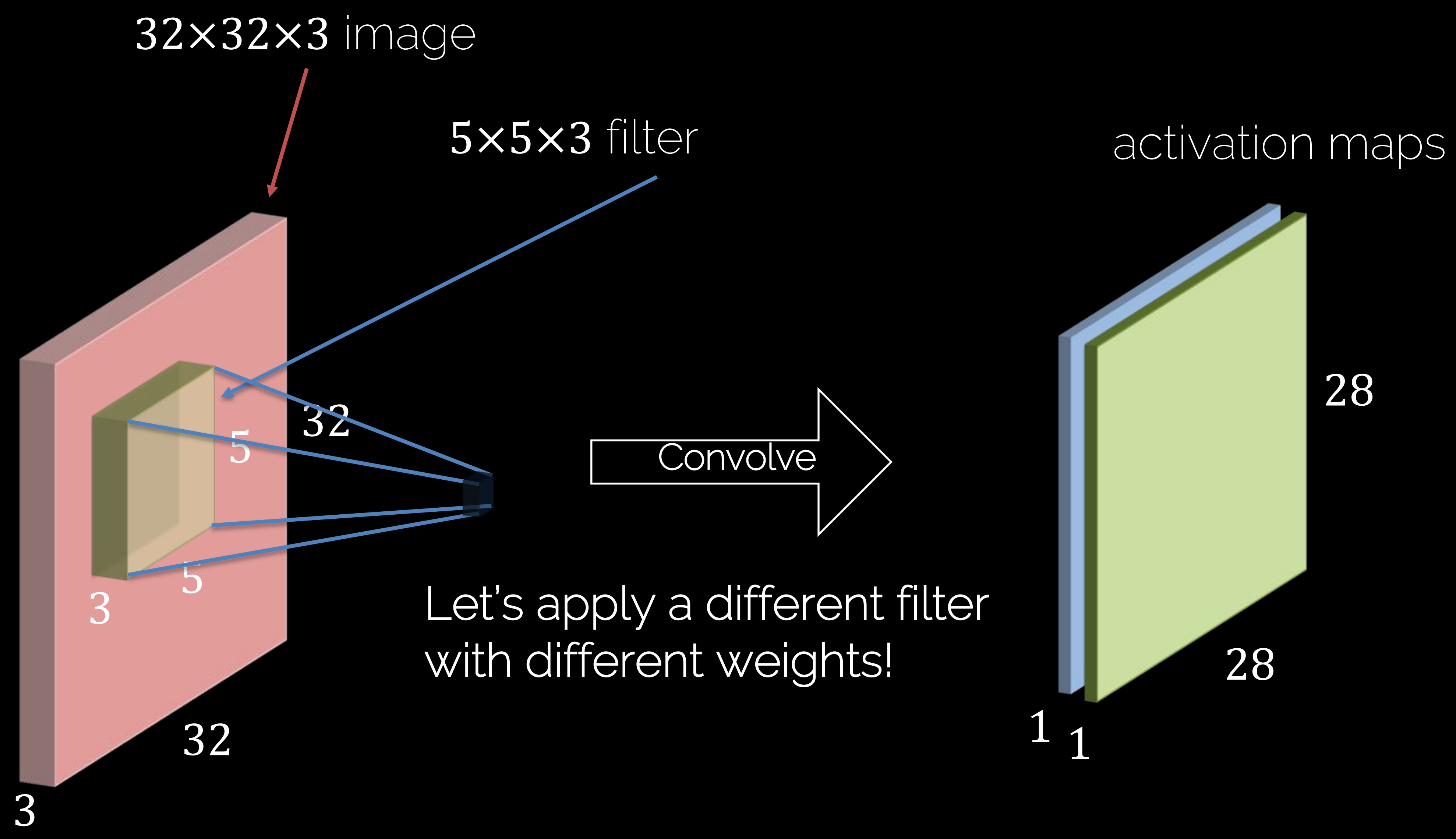
$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

LET'S LEARN THESE FILTERS!

Convolutions on RGB Images

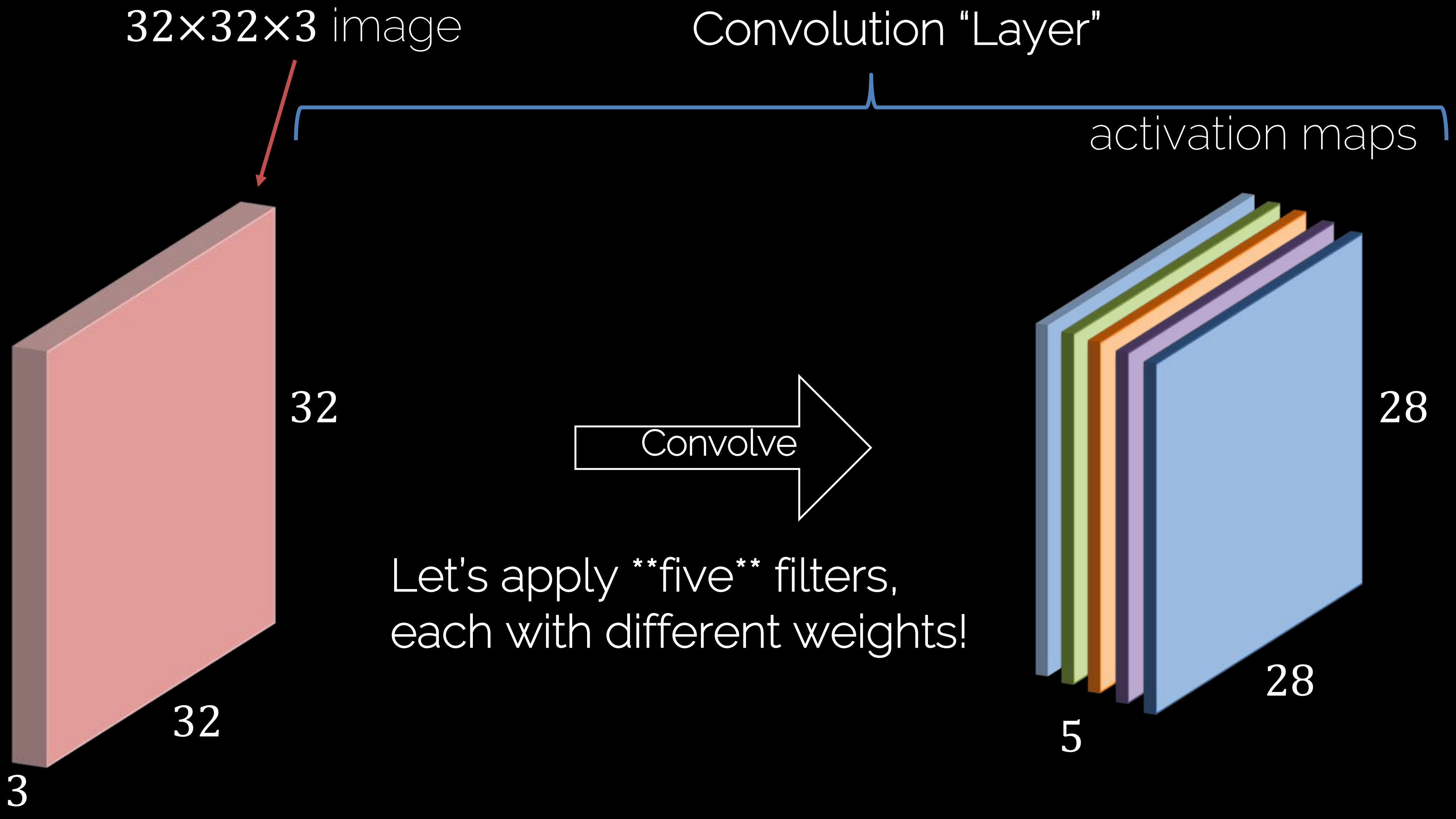


Convolution Layer

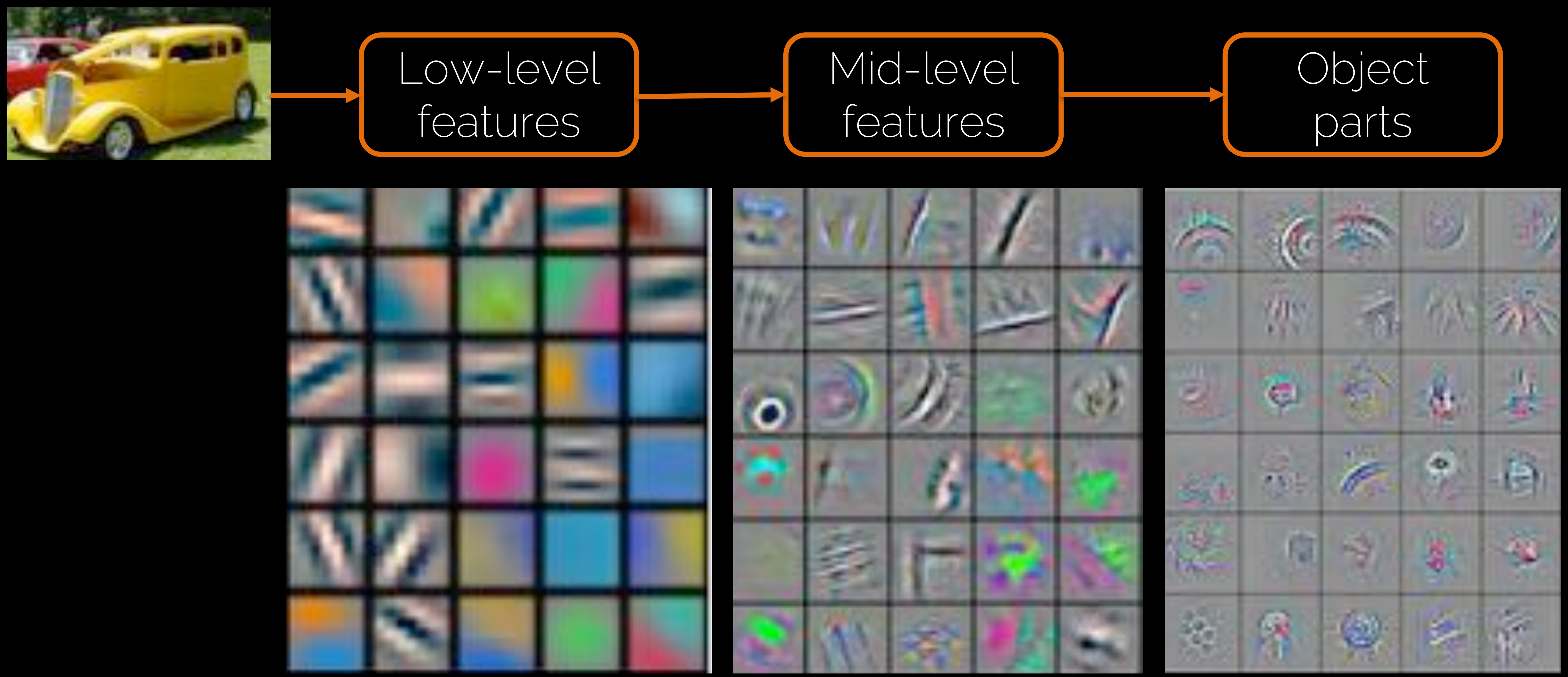


Let's apply a different filter with different weights!

Convolution Layer



Visualizing a CNN



What made Deep Learning possible?

Big Data

Models know where to learn from

Hardware

Models are trainable

Deep

Models are complex

Big data

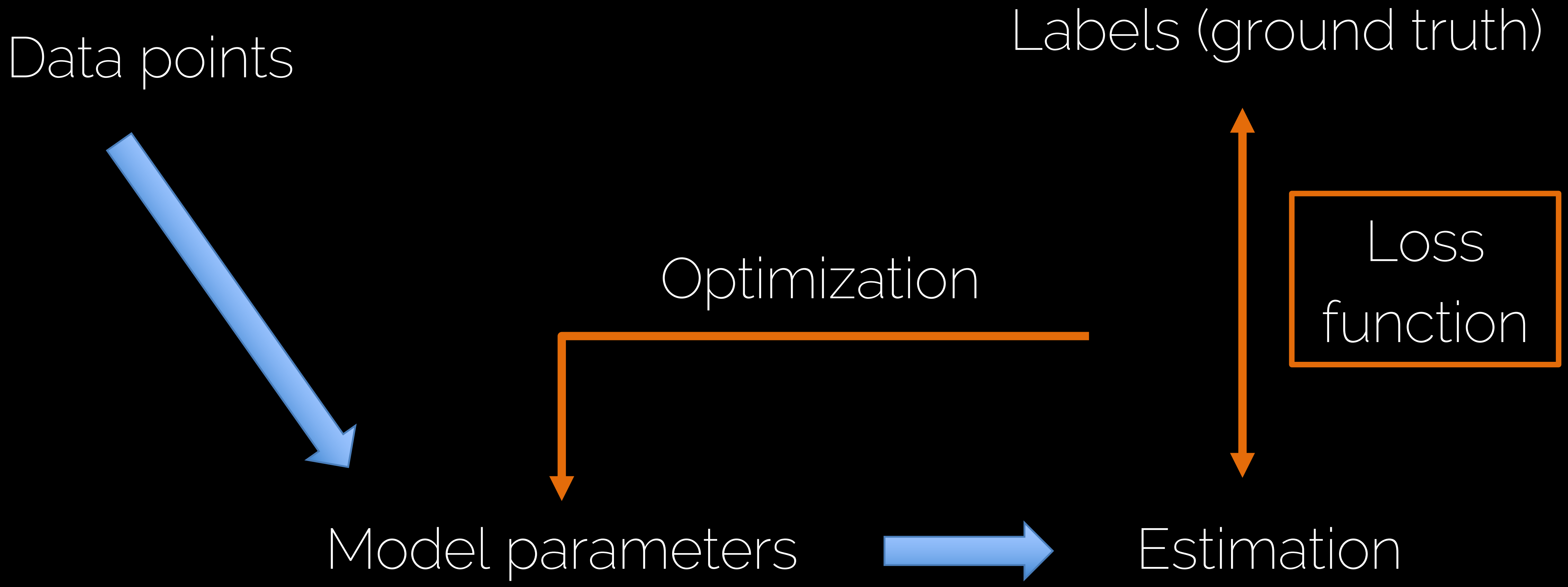
ImageNet: Goal 10.000 images per 100.000 words

Deep Learning: what is it good at?

Input A	→	Response B
English sentence	Machine translation	French sentence
Picture	Face recognition	Photo tagging
Audio clip	Speech recognition	Transcript

Supervised learning

How to obtain the model?

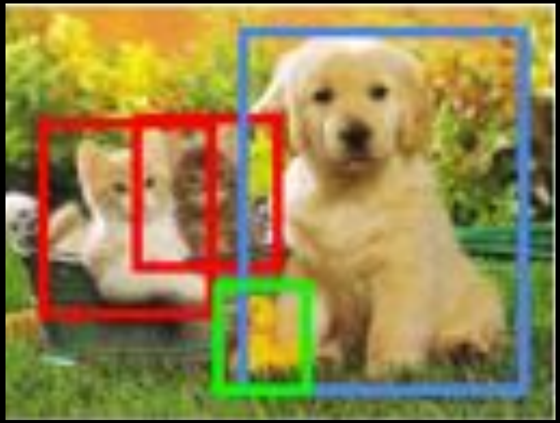


Deep Learning for Image Classification

Classification

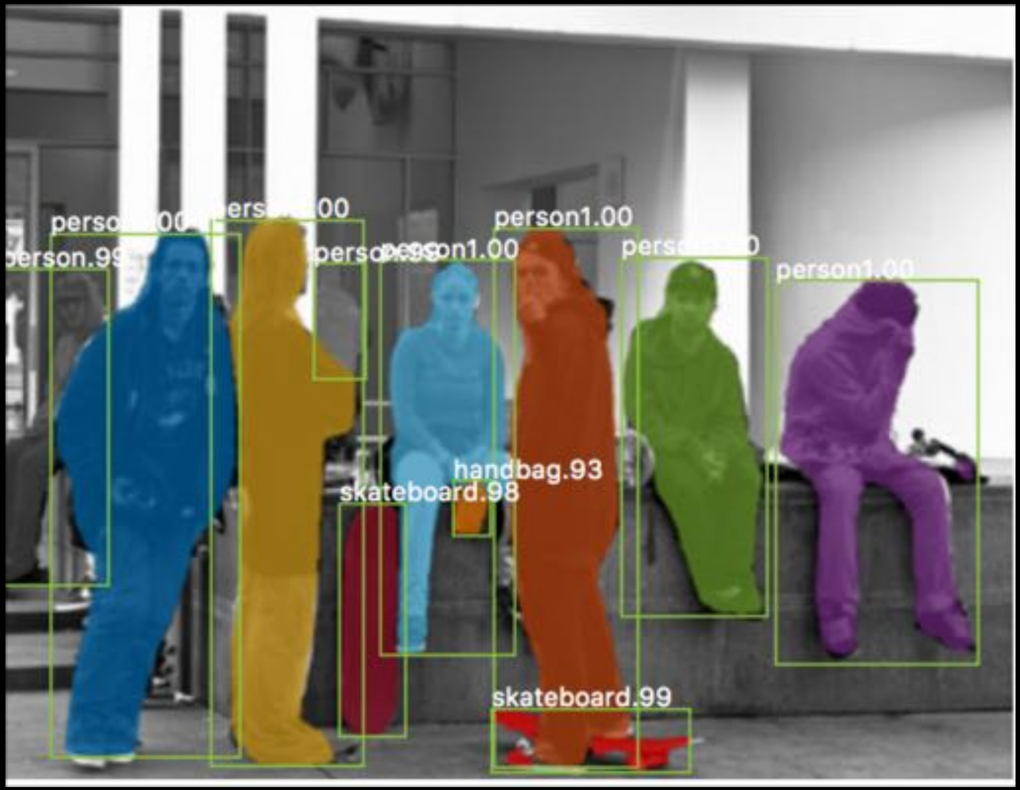
Classification
+
Localization

Object
detection



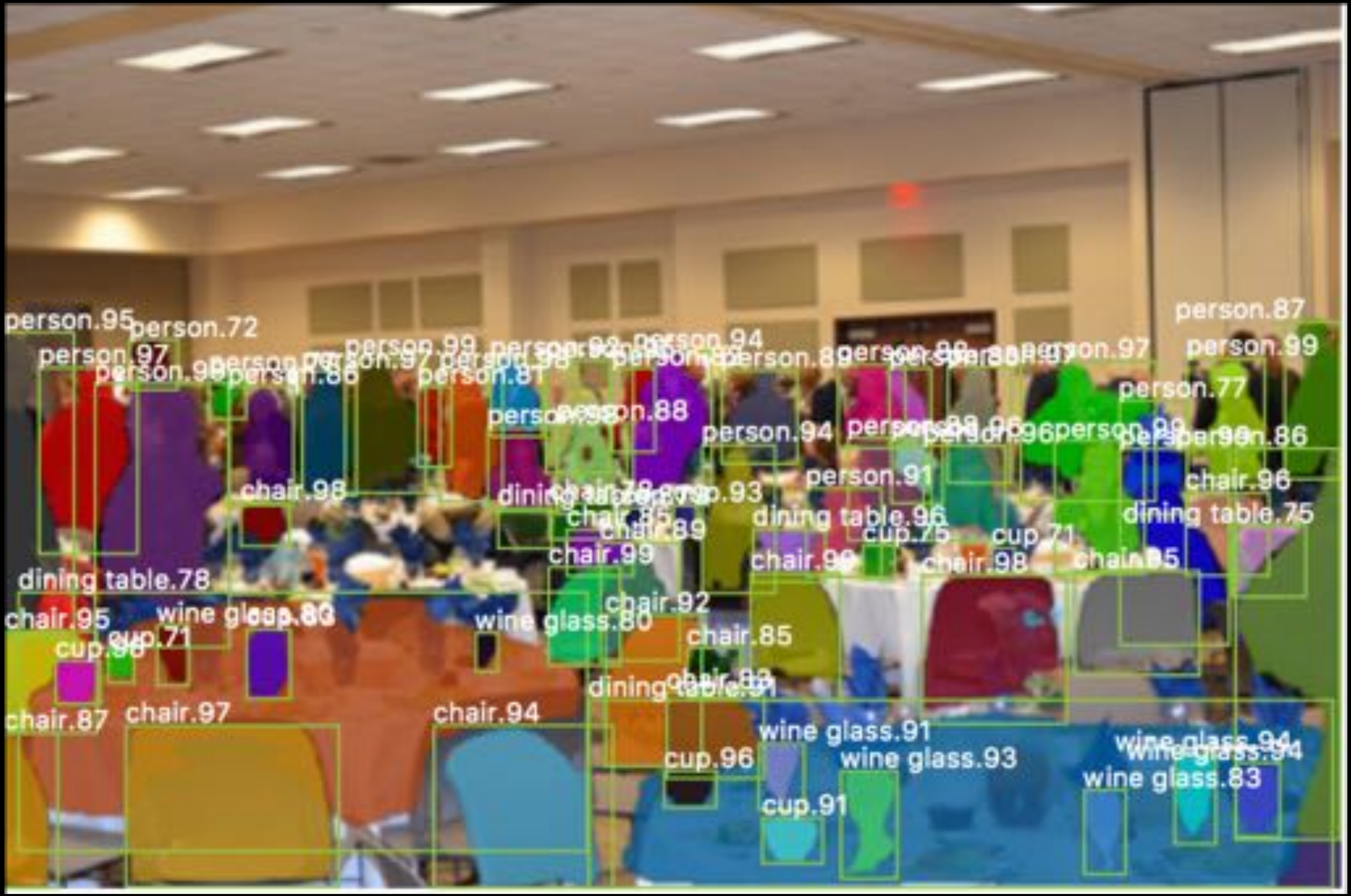
Instance
segmentation

Instance segmentation with Mask-RCNN



K. He, G. Gkioxari, P. Dollar, R. Girshick. Mask R-CNN. ICCV 2017.

Instance segmentation with Mask-RCNN



K. He, G. Gkioxari, P. Dollar, R. Girshick. Mask R-CNN. ICCV 2017.

Manipulating images

Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV, 2017.

Manipulating images

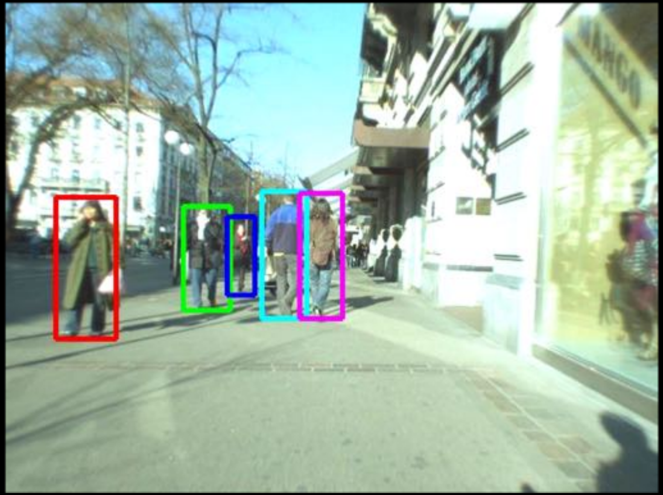
Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", ICCV, 2017.

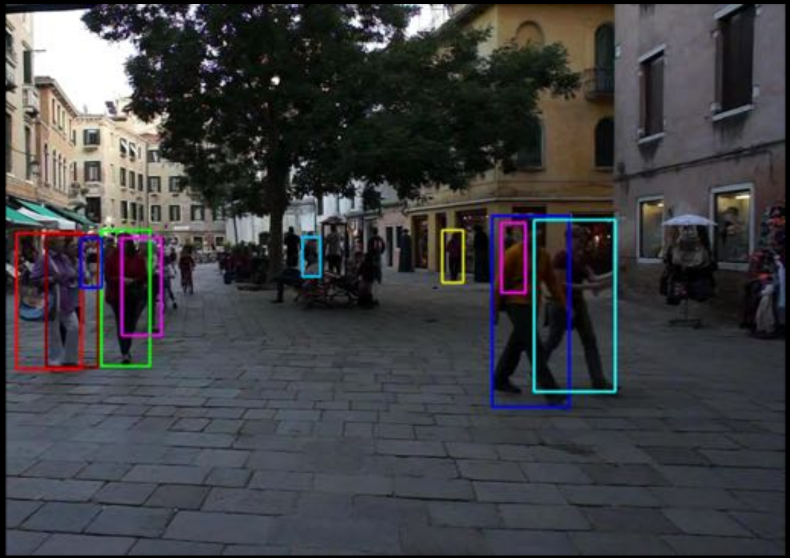
The world is dynamic!

- Deep Learning pushed single-image analysis to a point where results are usable in real-world scenarios
- The world is not static

What we still need in ML: good memory models

Multiple object tracking





Goal: detect and track all objects in a scene

Video object segmentation

Video super resolution

M. Chu, Y. Xie, L. Leal-Taixé and N. Thuerey. „Temporally Coherent GANs for Video Super-Resolution (TecoGAN)"

Video super resolution

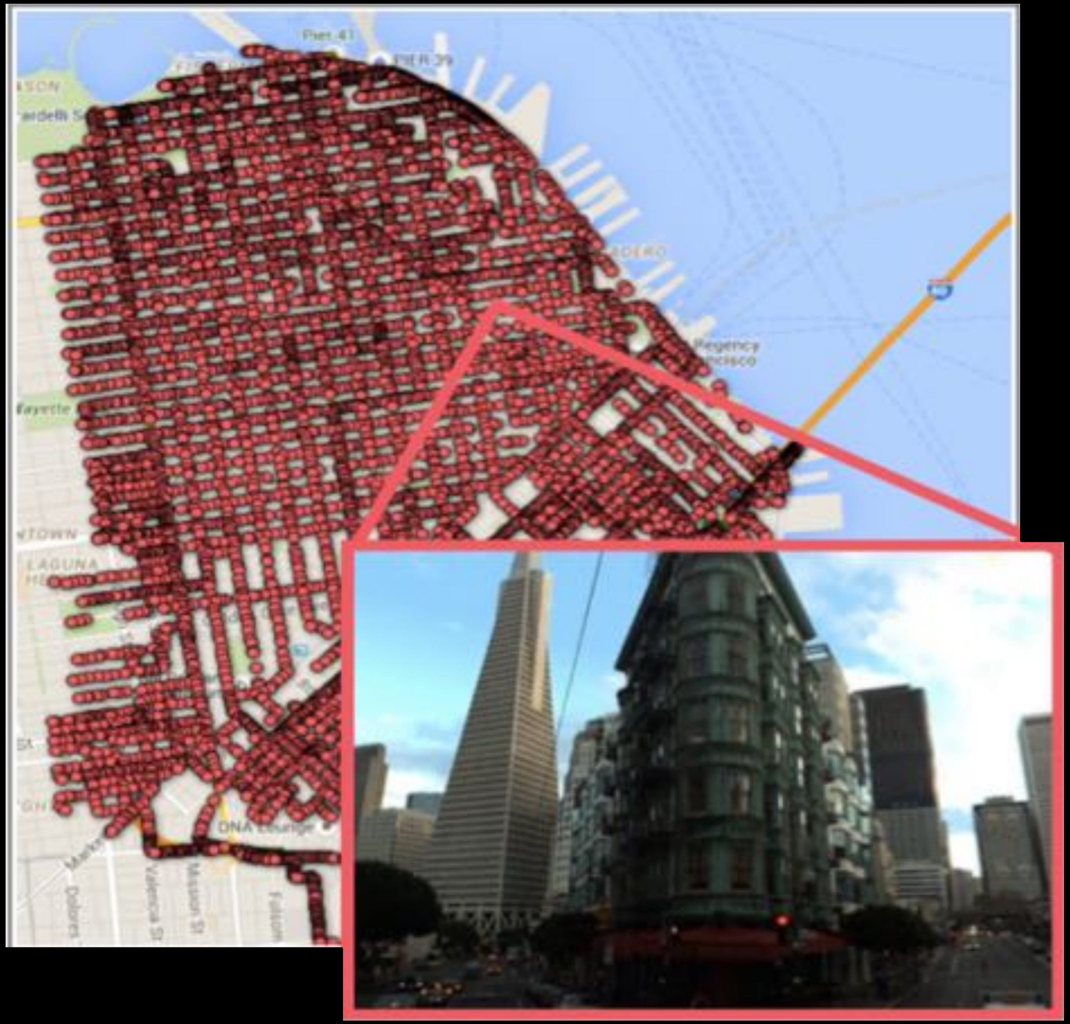
M. Chu, Y. Xie, L. Leal-Taixé and N. Thuerey. „Temporally Coherent GANs for Video Super-Resolution (TecoGAN)“

TecoGAN results

M. Chu, Y. Xie, L. Leal-Taixé and N. Thuerey. „Temporally Coherent GANs for Video Super-Resolution (TecoGAN)“

Visual localization

Map

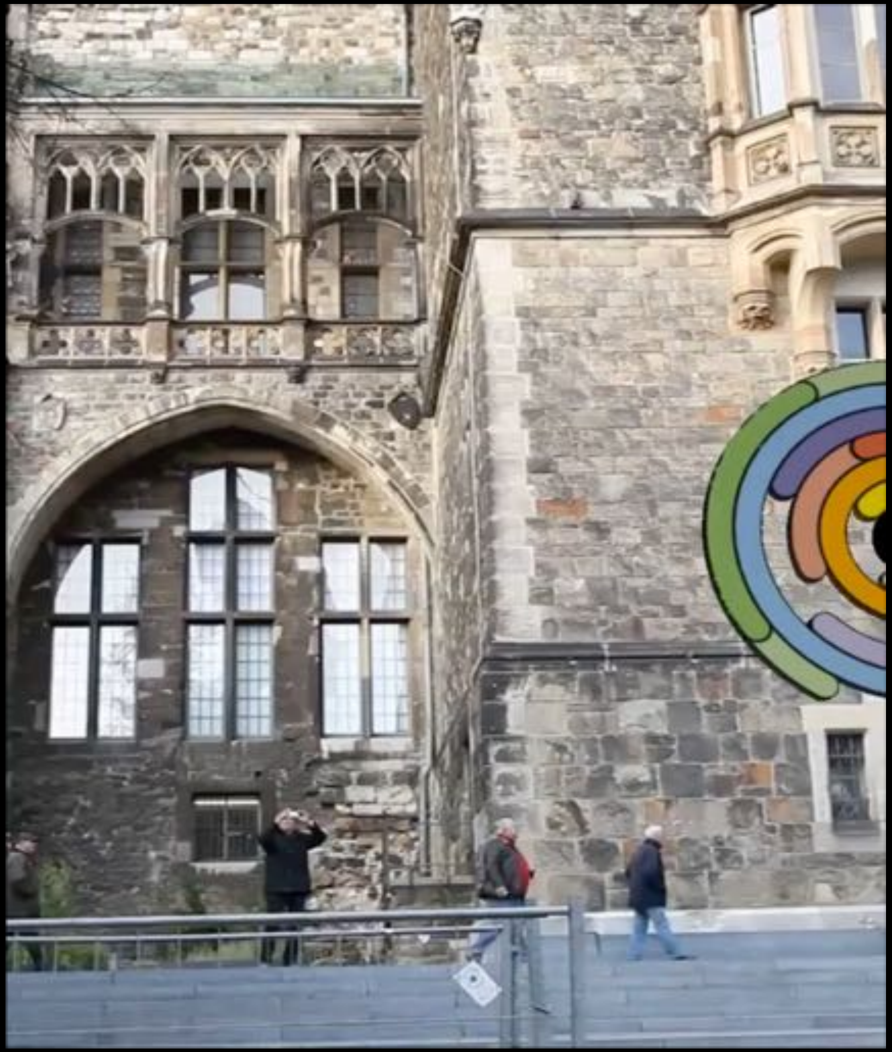


Obtain the camera pose with respect to the map

Photo

Visual Localization: applications

- Robot navigation
- Augmented reality



[Middelberg, Sattler, Untzelmann, Kobbelt, Scalable 6-DOF Localization on Mobile Devices. ECCV 2014]

The challenge: Big data

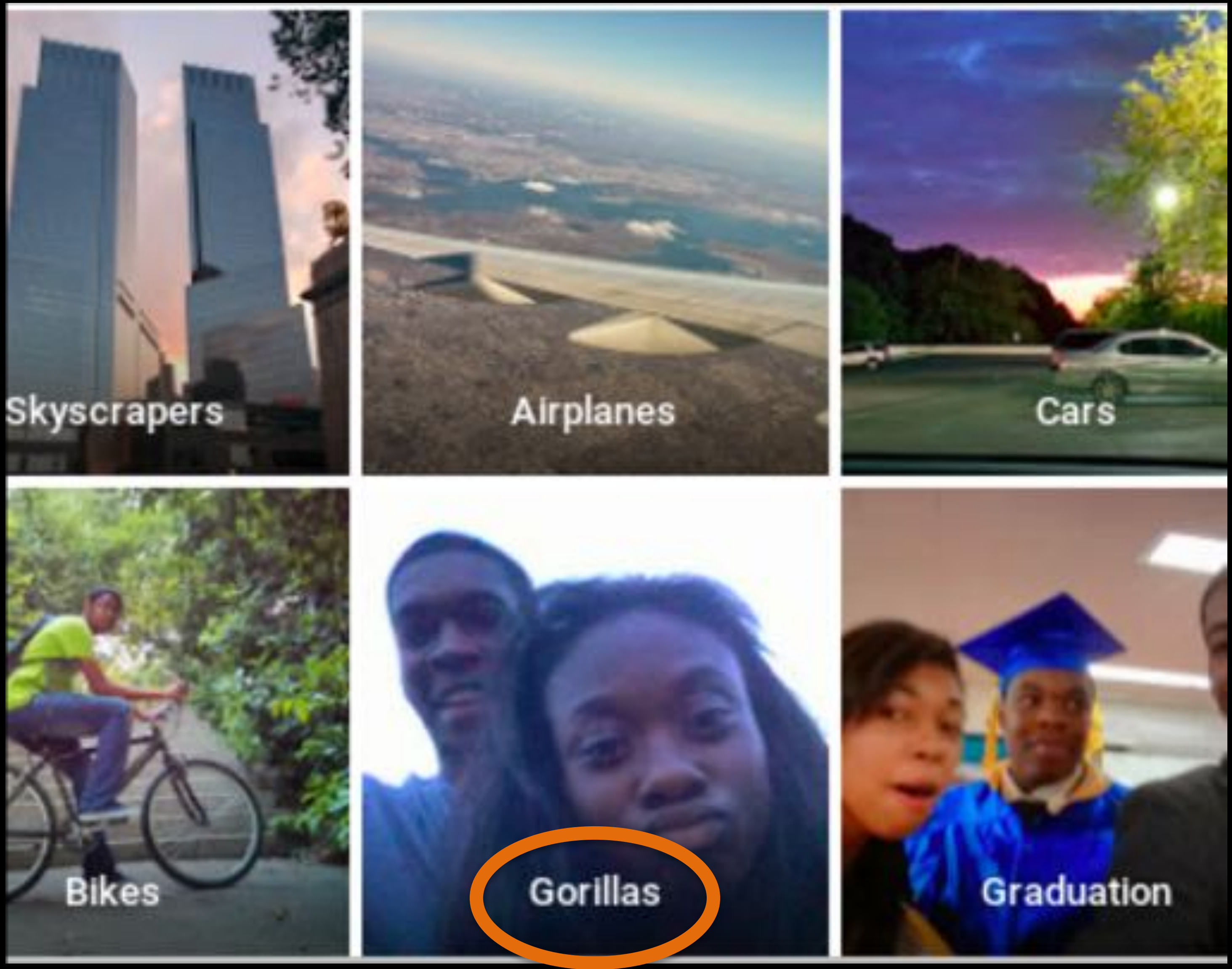
- We need data, lots of data!
- We might not have the luxury of data for some applications such as medical diagnosis
- Data is biased

Big data

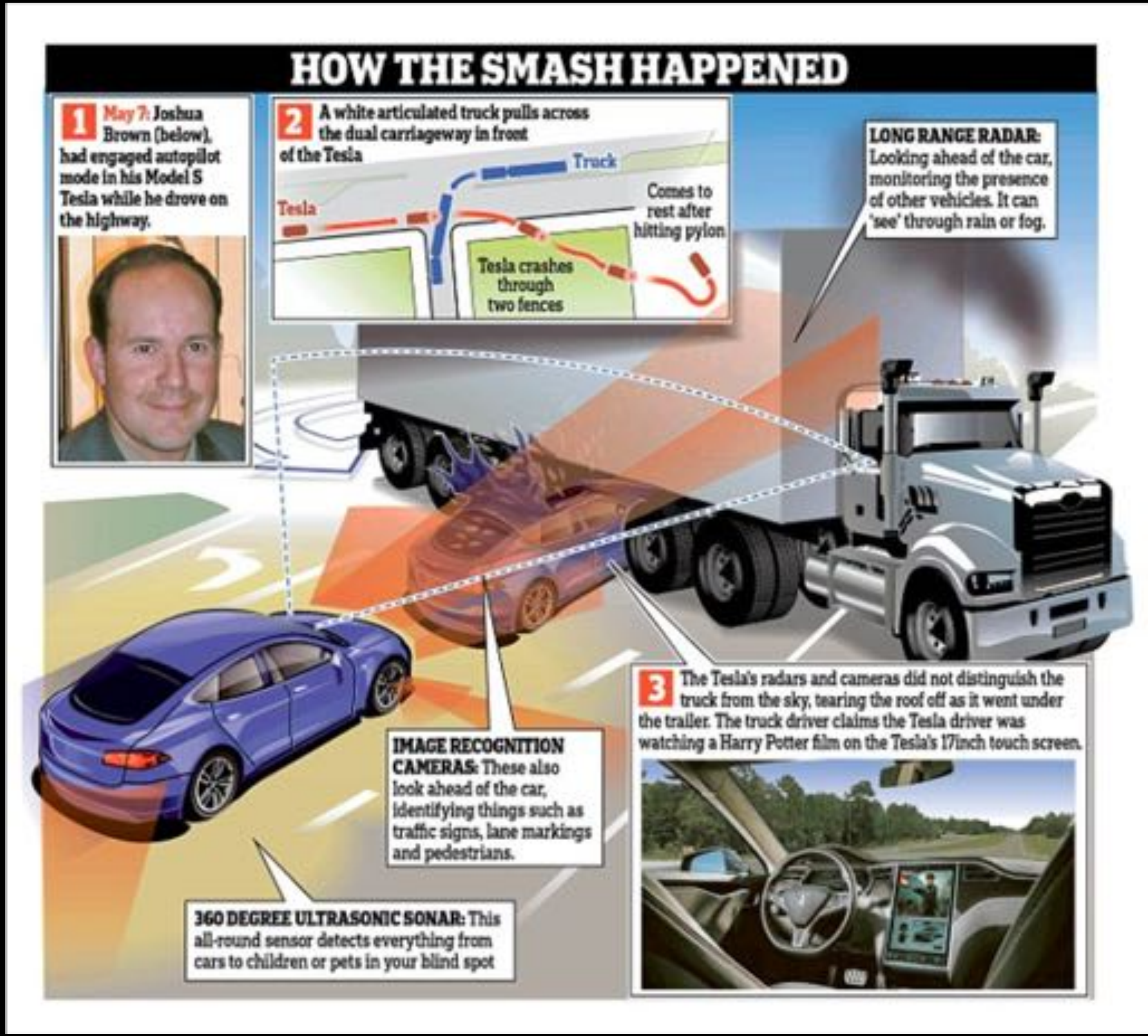
Data bias

- Increase diversity in the data

- Increase diversity in the AI community which is building the algorithms



Data bias



HOW THE SMASH HAPPENED
1 May 7: Joshua Brown (below), had engaged autopilot mode in his Model S Tesla while he drove on the highway.
2 A white articulated truck pulls across the dual carriageway in front of the Tesla
Truck
Tesla
Comes to rest after hitting pylon
Tesla crashes through two fences
LONG RANGE RADAR: Looking ahead of the car, monitoring the presence of other vehicles. It can 'see' through rain or fog.
3 The Tesla's radars and cameras did not distinguish the truck from the sky, tearing the roof off as it went under the trailer. The truck driver claims the Tesla driver was watching a Harry Potter film on the Tesla's 17inch touch screen.
IMAGE RECOGNITION CAMERAS: These also look ahead of the car, identifying things such as traffic signs, lane markings and pedestrians.
360 DEGREE ULTRASONIC SONAR: This all-round sensor detects everything from cars to children or pets in your blind spot

The generalization problem

- Neural networks are GREAT at finding patterns in data they have seen, but not so great at **generalizing** to new scenarios

True intelligence is still far away

Thank you

Prof. Dr. Laura Leal-Taixé

www.dvl.in.tum.de
https://dvl.in.tum.de

If you have images, contact us!

Technische Universität München